畳み込みニューラルネットワークを用いた
ひらがなのくずし字認識
Historical Cursive(Kuzushiji) Hiragana Character Recognition using Convolutional Neural Networks

学籍番号：201621646
氏名：原　皓
Kou GEN

歴史の研究には古典籍の解読が不可欠である．近代以前における日本語の古典籍の中で用いられた書体は、字が崩れている「くずし字」が多く、かつ複数の文字が連続して書かれているため、一般人にとって認識は難しい．近年、深層学習の発展とともに、文字認識の分野にも新たな進歩がもたらされた．その中で、畳み込みニューラルネットワークは画像認識に特化した手法として、次々と成果が現れている．本研究で、日本語古典籍中のひらがなのくずし字を、畳み込みニューラルネットワークを用いて認識する．

研究で、人文学オープンデータ共同利用センターが公開した日本古典籍字形データセットの中の字形画像データを用いた．その中で、ひらがなのくずし字は、各文字種のデータ数が不均衡であった．データ数の多い文字種は一万件以上に対し、データ数の少ない文字種はたった数個しかなく、またすべてのひらがなのくずし字において、濁点や半濁点のあるデータ数は少ないという特徴もあった．そのため、画像認識におけるデータの不均衡問題を改善するために、一般的な濁点と半濁点のあるひらがなも含め全 73 種のひらがなのくずし字を直接判別する手法以外、並列手法、順次手法 1 と順次手法 2 の新しい手法を三つ提案した．①並列手法では、濁点・半濁点の有無を無視し、形が違う 48 種のひらがなの判別と、濁点・半濁点の有無の判別を並列に行い、最後に両方の結果を合わせる．②順次手法 1 では、濁点・半濁点の有無を無視し、形が違う 48 種のひらがなを判別した後、出力した文字コードに基づいて、濁点と半濁点の有無を判別する．③順次手法 2 では、半濁点のあるひらがなだけを除いて、68 種のひらがなを判別した後、半濁点の有無を判別する．

実験の結果、ひらがなの判別は 73 種から 48 種に減少させても、正答率に影響はなかった．濁点と半濁点なし、濁点あり、半濁点ありの 3 分類から、①濁点なしと濁点あり、②濁点ありと半濁点ありの 2 分類にすることによって、判別精度が 96.52%と 96.99%から 98.37%と 98.34%に上がった．全体的に、データ数が充分かつ各クラスのデータのバランスが取れた状況ならば、全 73 種のひらがなのくずし字を直接判別するのが一番良い方法と推測できる．データセットの状況、すなわち半濁点のあるひらがなのくずし字のデータ数が足りない状況から、提案した新しい 3 種の手法の正答率を確率的に算出すると、並列手法の正答率は 94.71%、順次手法 1 の正答率は 96.86%、そして順次手法 2 最も高い正答率 97.93%を得た．

研究指導教員：長谷川　秀彦
副研究指導教員：手塚　太郎